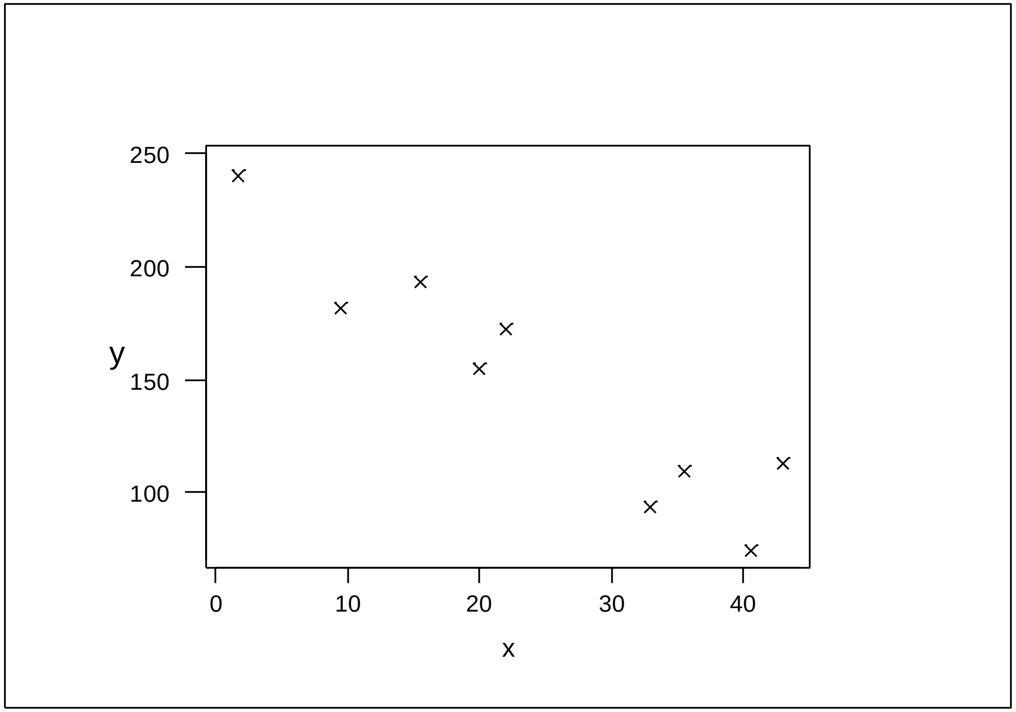
Stats for Engineers Lecture 10

Recap: Linear regression

We measure a response variable y at various values of a controlled variable x



Linear regression: fitting a straight line to the *mean* value of y as a function of x

Equation of the fitted line is $\hat{y} = \hat{a} + \hat{b}x$

Least-squares estimates $\hat{a}$ and $\hat{b}$:

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

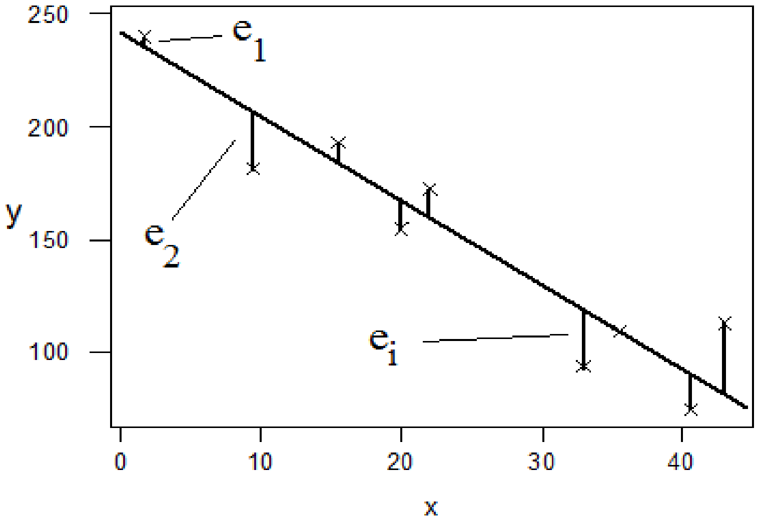
Sample means

$$S_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = \sum_i (x_i - \bar{x})^2$$

$$S_{xy} = \sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Quantifying the goodness of the fit

Estimating σ^2: variance of *y* about the fitted line



$$\hat{\sigma}^2 = \frac{1}{n-2}\sum \hat{e}_i^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2$$

$$= \frac{S_{yy} - \hat{b}S_{xy}}{n-2}$$

Residual sum of squares

Predictions

For given x of interest, what is mean y?

Predicted mean value: $\hat{y} = \hat{a} + \hat{b}x$.

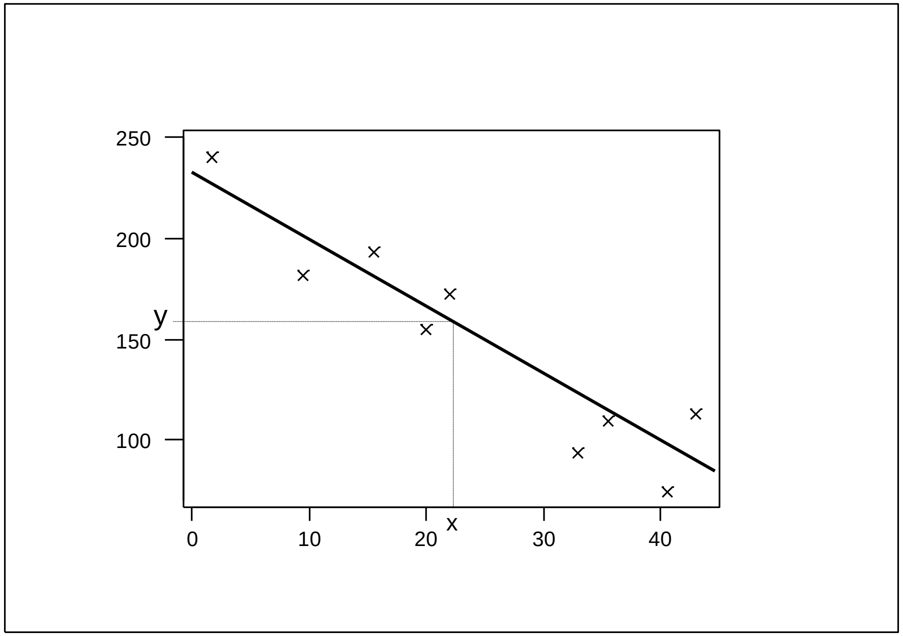
What is the error bar?

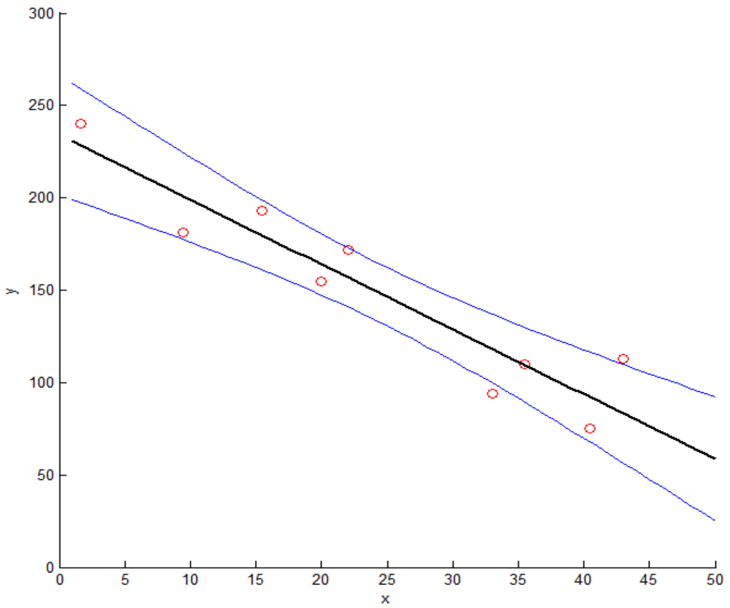
It can be shown that

$$\begin{aligned}\mathrm{var}(\hat{y}|x) &= \mathrm{var}\left(\hat{a} + \hat{b}x\right)\\ &= \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\end{aligned}$$

Confidence interval for mean *y* at given *x*

$$\hat{y} \pm t_{n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}$$





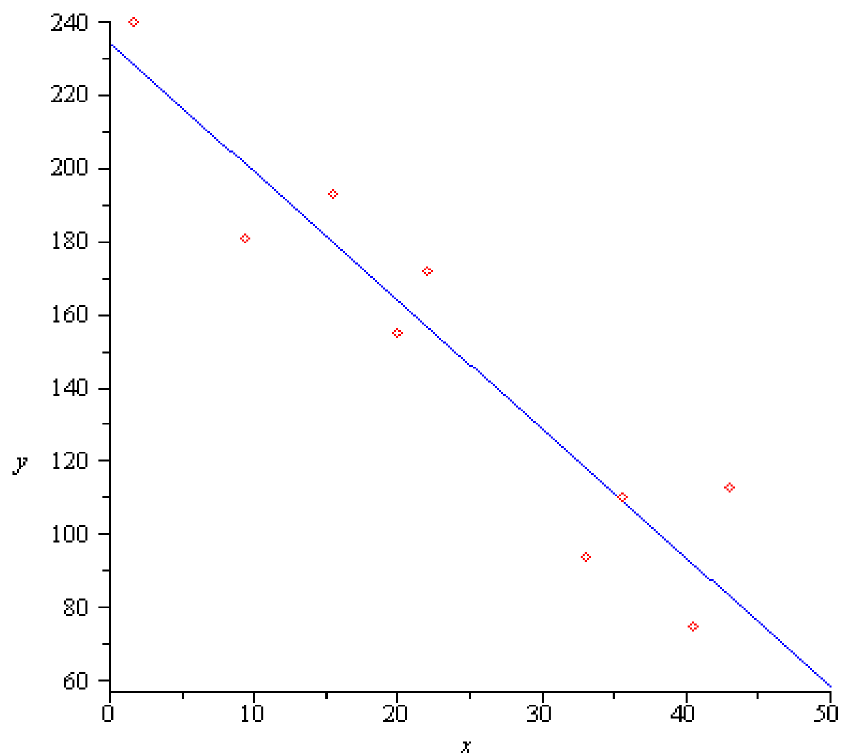
***Example*:**

The data *y* has been observed for various values of *x*, as follows:

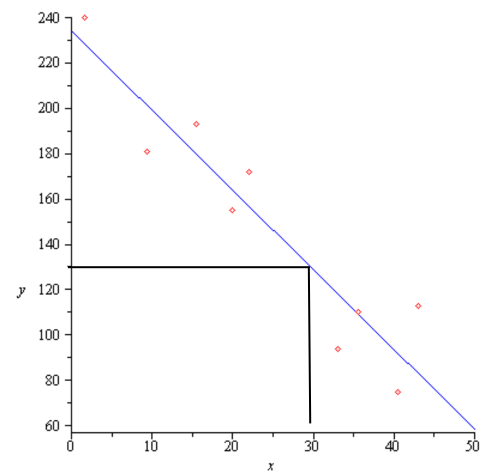
y	240	181	193	155	172	110	113	75	94
x	1.6	9.4	15.5	20.0	22.0	35.5	43.0	40.5	33.0

Fit the simple linear regression model using least squares.

$$\hat{y} = 234.1 - 3.509x$$



Example: Using the previous data, what is the mean value of y at $x = 30$ and the 95% confidence interval?



Recall fit was $y = 234.1 - 3.509x$

$$x = 30 \Rightarrow \hat{y} = 234 - 3.509 \times 30 = 128.8$$

Confidence interval is $\hat{y} \pm t_{n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}$, $n = 9$

Need

$$\widehat{\sigma}^2 = \frac{S_{yy} - \hat{b}S_{xy}}{n-2} = 398.28$$

95% confidence $\Rightarrow t_{n-2} = t_7$ for Q=0.975 $\quad \Rightarrow t_{n-2} = t_7 = 2.3646.$

Hence confidence interval for mean y is

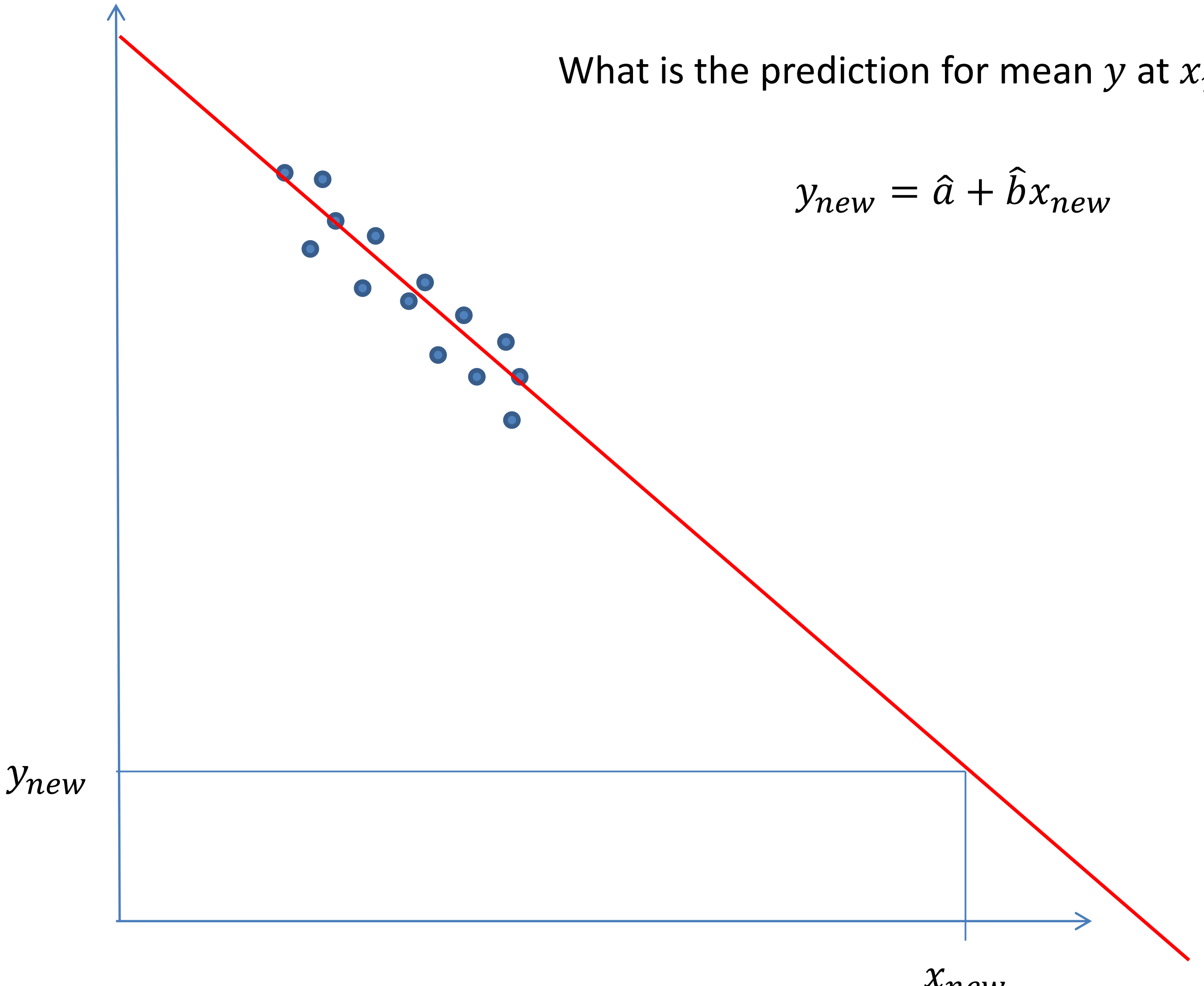
$$\hat{y} \pm t_7\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(30 - \bar{x}\,)^2}{S_{xx}}\right)} = 128.8 \pm 2.3646\sqrt{398.28\left(\frac{1}{9} + \frac{\left(30 - \frac{220.5}{9}\right)^2}{1651.42}\right)}$$

$$\approx 129 \pm 17$$

Extrapolation: *predictions outside the range of the original data*

What is the prediction for mean y at x_{new}?

$$y_{new} = \hat{a} + \hat{b}x_{new}$$

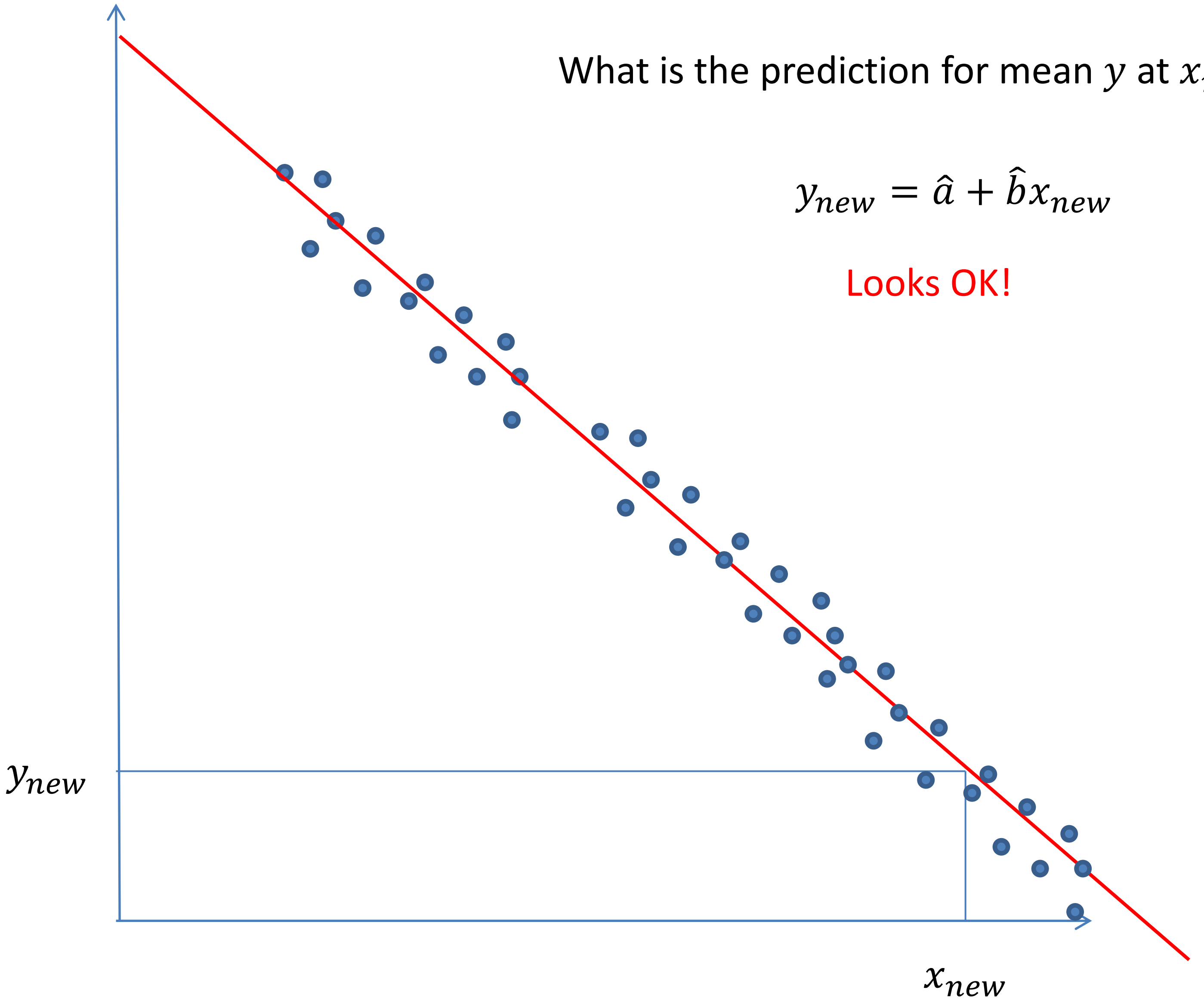


Extrapolation: *predictions outside the range of the original data*

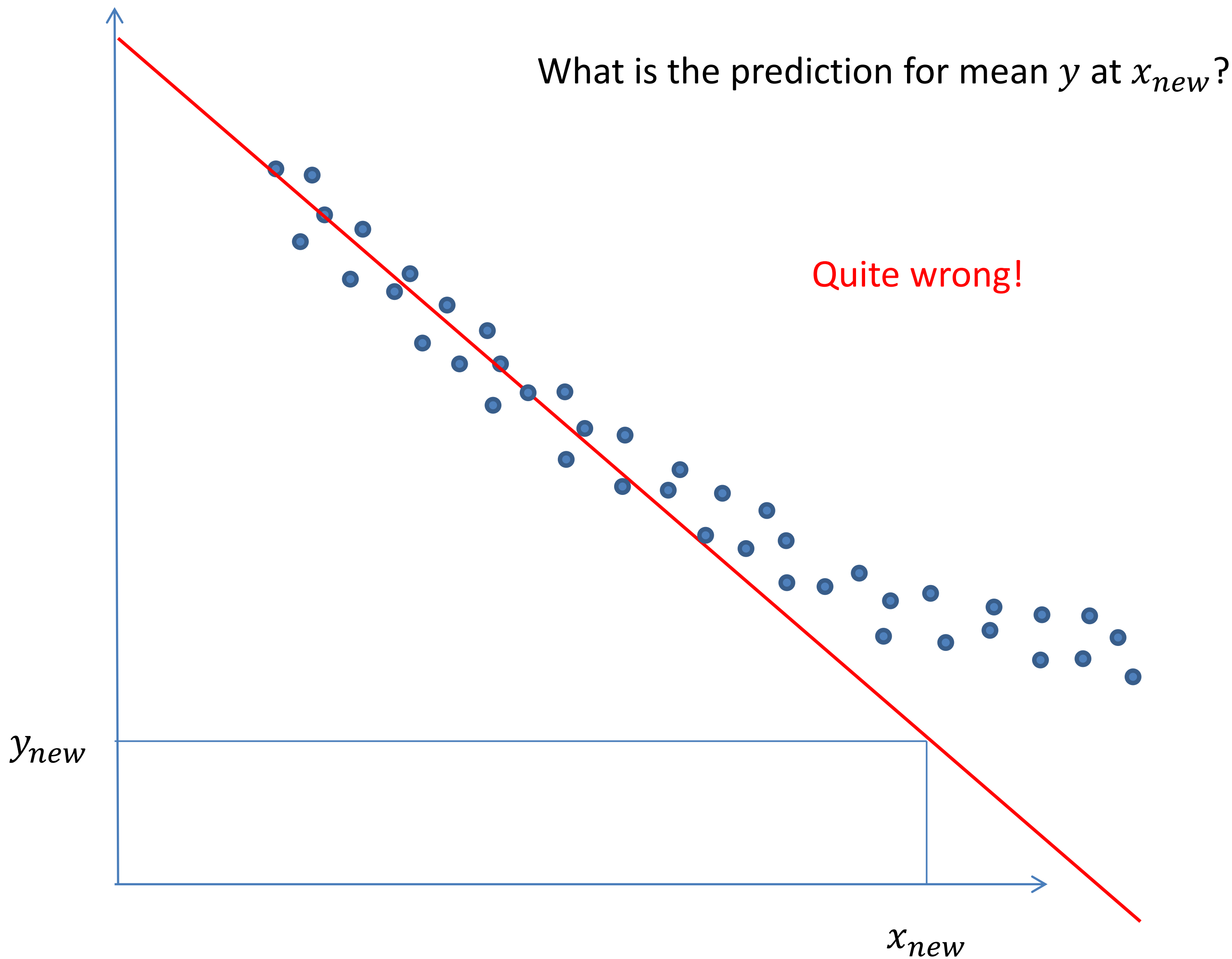
What is the prediction for mean y at x_{new}?

$$y_{new} = \hat{a} + \hat{b}x_{new}$$

Looks OK!



Extrapolation: *predictions outside the range of the original data*



Extrapolation is often unreliable unless you are sure straight line is a good model

We previously calculated the confidence interval for the *mean:* if we average over many data samples of y at x, this tells us the interval we expect the *average* y to lie in.

What about the distribution of future data points themselves?

Confidence interval for a prediction

Two effects:

- Variance on our estimate of *mean* y at x $\quad \hat{\sigma}^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)$
- Variance of individual points about the mean $\quad \hat{\sigma}^2$

$\Rightarrow$ Confidence interval for a single response (measurement of y at x_0) is

$$y(x_0)=\hat{a}+\hat{b}x_0 \pm t_{n-2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

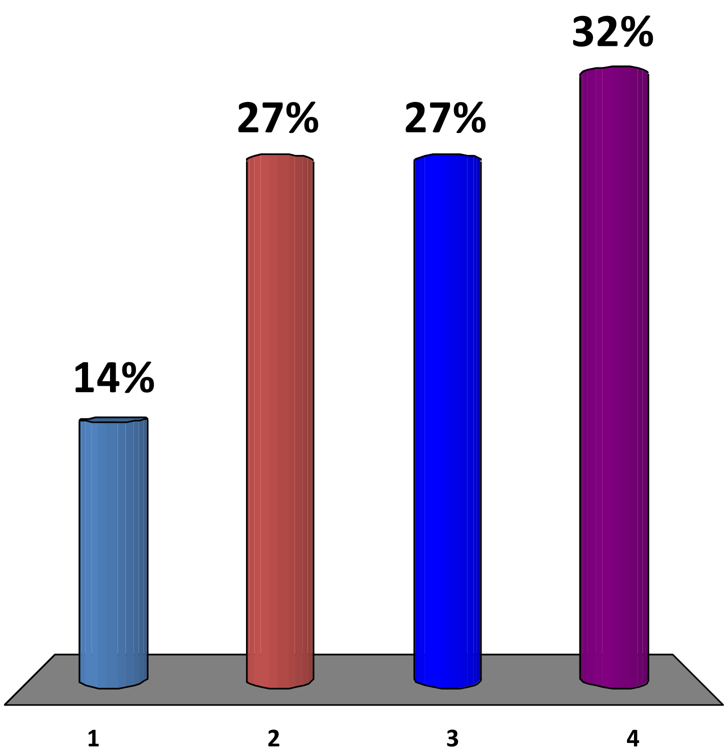
Example: Using the previous data, what is the 95% confidence interval for a new measurement of y at $x=30$?

Answer

$$\hat{y}\pm t_7\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(30-\bar{x})^2}{S_{xx}}\right)}=129\pm 2.36\sqrt{398.28\left(1+\frac{1}{9}+\frac{\left(30-\frac{220.5}{9}\right)^2}{1651.42}\right)}\approx 129\pm 50$$

A linear regression line is fit to measured engine efficiency y as a function of external temperature T (in Celsius) at values $T = (0,5,10,15,20,25,30)$. Which of the following statements is most likely to be **incorrect**?

1. The confidence interval for a new measurement of y at $\mathrm{T} = 15$ is narrower than at $\mathrm{T} = 30$
2. Adding a new data at $T = 40$ would decrease the confidence interval width at $T = 25$
3. If T and y accurately have a linear regression model, adding more data points at $T = 0$ and $T = 30$ would be better than adding more at $T = 15$ and $T = 20$
4. The mean engine efficiency at T= -20 will lie within the 95% confidence interval at T=-20 roughly 95% of the time



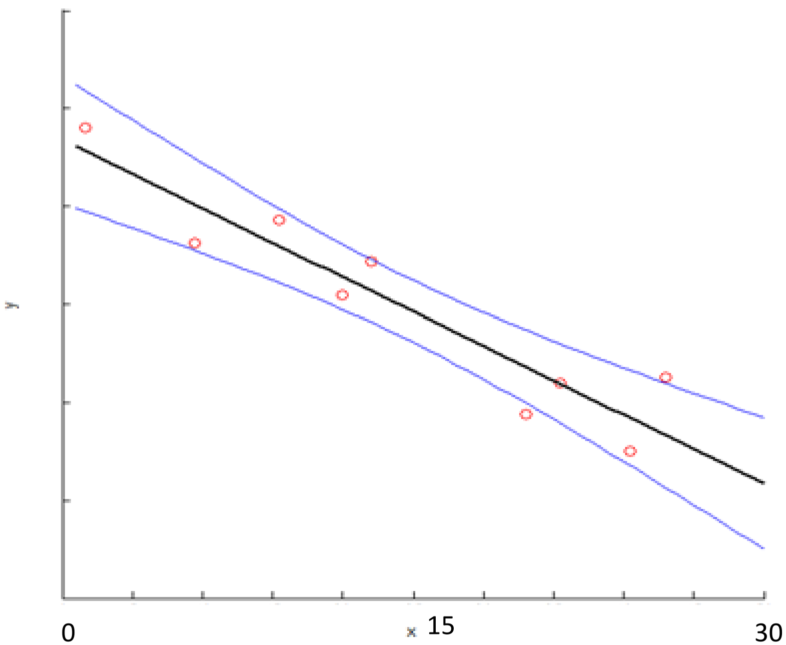
Answer

Confidence interval for mean *y* at given *x*

$$\hat{y} \pm t_{n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}$$

Confidence interval for a single response (measurement of y at x_0)

$$\hat{a} + \hat{b}x_0 \pm t_{n-2}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$



The confidence interval for a new measurement of y at $\mathrm{T} = 15$ is narrower than at $\mathrm{T} = 30$

- Confidence interval narrower in the middle ($x \sim \bar{x}$)

Adding a new data at $T = 40$ would decrease the confidence interval width at $T = 25$

- Adding new data decreases uncertainty in fit, so confidence intervals narrower (n larger)

If T and y accurately have a linear regression model, adding more data points at $T = 0$ and $T = 30$ would be better than adding more at $T = 15$ and $T = 20$

- If linear regression model accurate, get better handle on the slope by adding data at both ends (bigger S_{xx} ⇒smaller confidence interval)

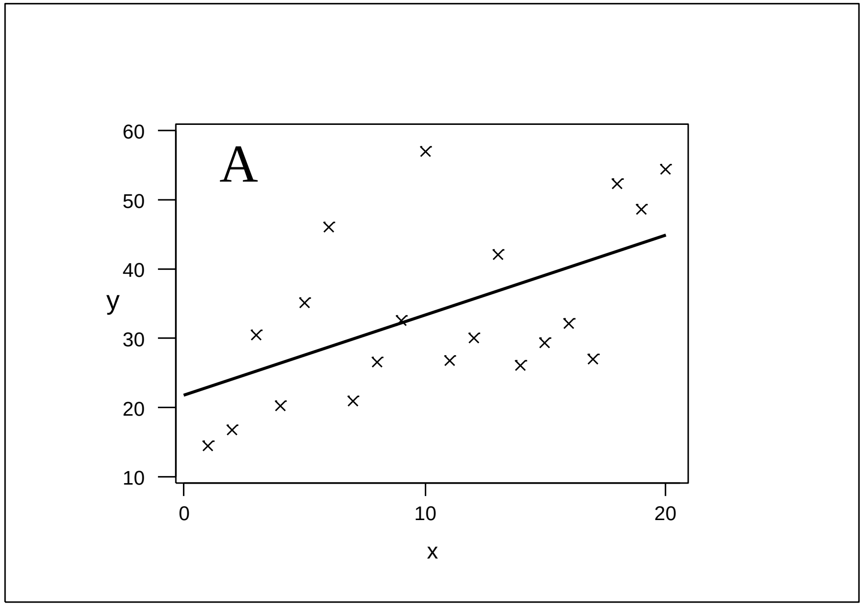
The mean engine efficiency at T= -20 will lie within the 95% confidence interval at T=-20 roughly 95% of the time

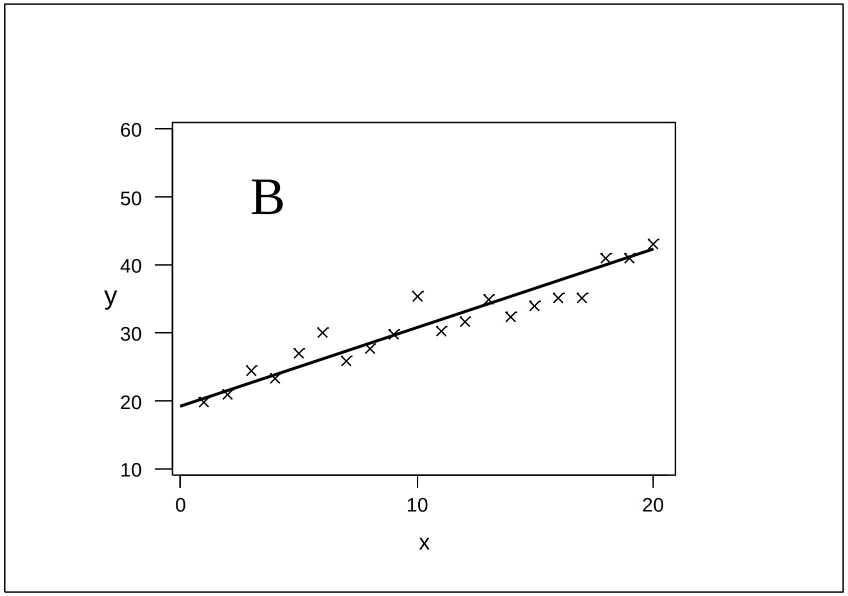
- Extrapolation often unreliable – e.g. linear model may well not hold at below-freezing temperatures. Confidence interval unreliable at T=-20.

Correlation

Regression tries to model the linear relation between mean *y* and *x*.

Correlation measures the strength of the linear association between *y* and *x*.





Weak correlation

Strong correlation

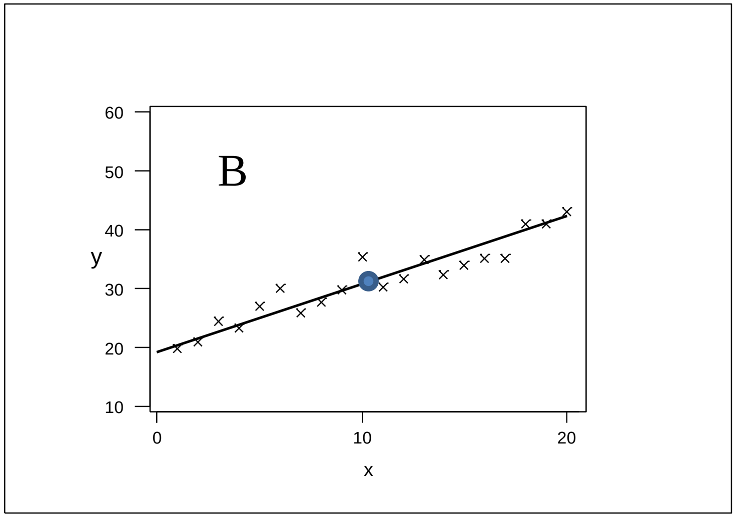
- same linear regression fit (with different confidence intervals)

If x and y are *positively* correlated:

- if x is high $(x > \bar{x})$ y is mostly high $(y > \bar{y})$

- if x is low $(x < \bar{x})$ y is mostly low $(y < \bar{y})$

$\Rightarrow$ on average $(x - \bar{x})(y - \bar{y})$ is *positive*



If x and y are *negatively* correlated:

- if x is high $(x > \bar{x})$ y is mostly low $(y < \bar{y})$

- if x is low $(x < \bar{x})$ y is mostly high $(y < \bar{y})$

$\Rightarrow$ on average $(x - \bar{x})(y - \bar{y})$ is *negative*

$\Rightarrow$ can use $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$ to quantify the correlation

More convenient if the result is independent of units (dimensionless number).

Define

$$\boxed{r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}}$$

Pearson product-moment.

If $y \to Ay$, then r is unchanged ($S_{xy} \to AS_{xy}, S_{yy} \to A^2 S_{yy}$)
Similarly for x - stretching plot does not affect r.

Range $-1 \leq r \leq 1$:

r = 1: there is a line with positive slope going through all the points;

r = -1: there is a line with negative slope going through all the points;

r = 0: there is no linear association between *y* and *x*.

***Example*: from the previous data: $S_{xy} = -5794, S_{xx} = 1651, S_{yy} = 23117$**

Hence

$$r = \frac{-5794}{\sqrt{1651 \times 23117}} \approx -0.94$$

Notes:

- magnitude of *r* measures how noisy the data is, but not the slope

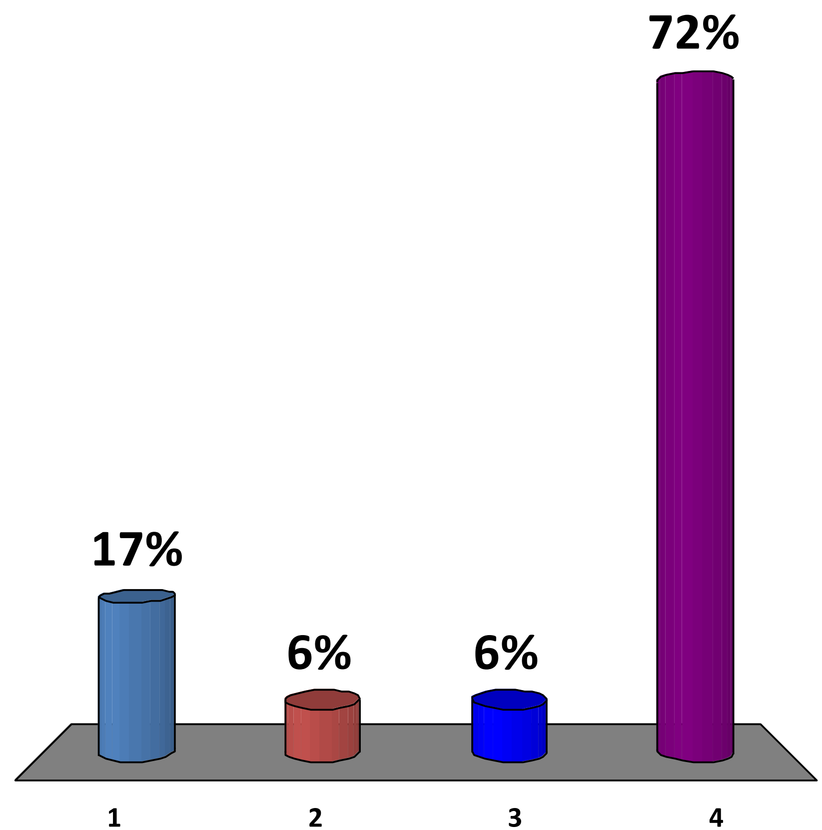
- finding $r = 0$ only means that there is no linear relationship, and does not imply the variables are independent

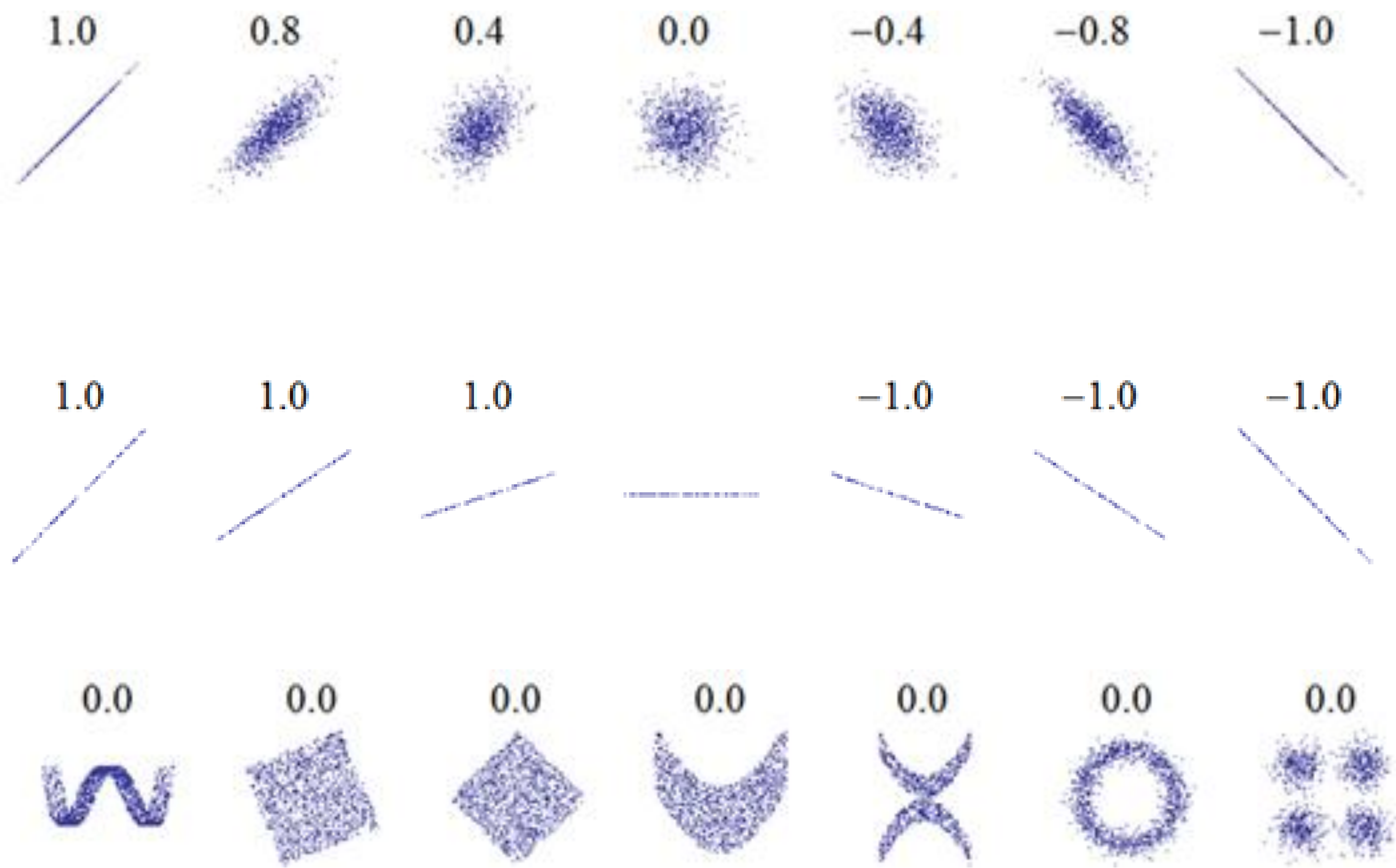
Question from Murphy et al.

Correlation

A researcher found that r = +0.92 between the high temperature of the day and the number of ice cream cones sold in Brighton. What does this information tell us?

1. Higher temperatures cause people to buy more ice cream.
2. Buying ice cream causes the temperature to go up.
3. Some extraneous variable causes both high temperatures and high ice cream sales
4. Temperature and ice cream sales have a strong positive linear relationship.




1.0
0.8
0.4
0.0
−0.4
−0.8
−1.0
1.0
1.0
1.0
−1.0
−1.0
−1.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0

Error on the estimated correlation coefficient?

- not easy; possibilities include subdividing the points and assessing the spread in *r* values.

Causation? $r \neq 0$ does *not* imply that changes in *x* *cause* changes in *y* - additional types of evidence are needed to see if that is true.

Correlation r

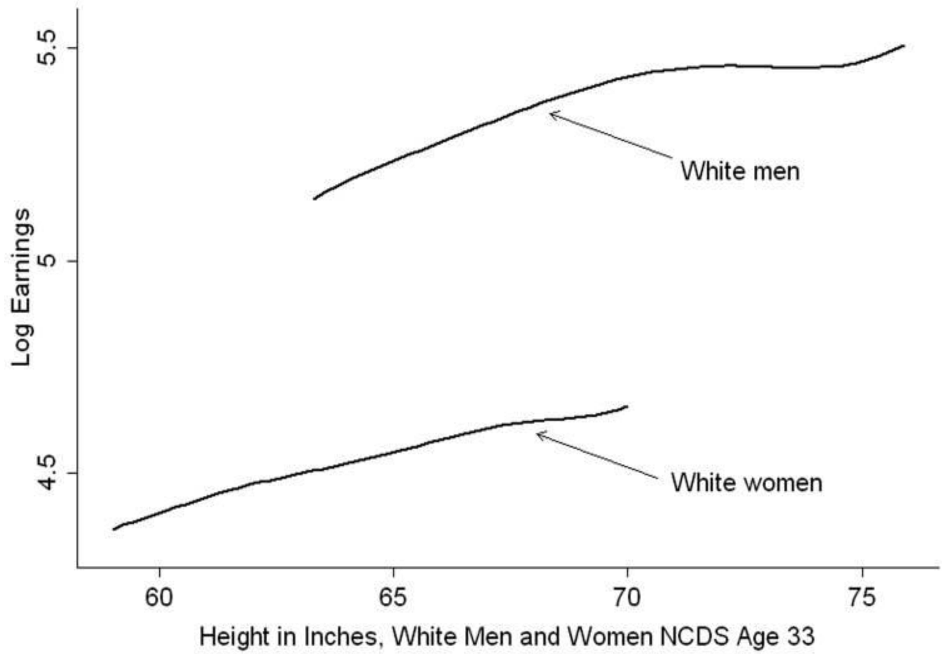
error

Log earnings and height

	Men		Women	
Dependent variable:	**Height coefficient**	**Number of observations**	**Height coefficient**	**Number of observations**
		NCDS		
Log weekly gross earnings	0.026	4,927	0.024	5,033
	(0.004)		(0.007)	
Log average hourly gross earnings	0.023	4,860	0.019	4,995
	(0.004)		(0.005)	
		BCS		
Log weekly gross earnings	0.014	2,265	0.029	2,136
	(0.003)		(0.006)	
Log average hourly gross earnings	0.010	2,253	0.015	2,127
	(0.003)		(0.004)	
		PSID		
Log weekly earnings	0.023	23,465	0.014	21,271
	(0.004)		(0.006)	
Log average hourly earnings	0.019	23,465	0.012	21,271
	(0.004)		(0.003)	

Notes. OLS regression coefficients reported for height in inches, with standard errors in parentheses. The NCDS and PSID regressions use multiple observations per person, and unobservables are clustered at the individual level. The NCDS and BCS samples are restricted to those for whom we have test scores at ages 7 and 11 (NCDS), or 5 and 10 (BCS). The PSID sample consists of white household heads or wives between the ages of 25 and 60, inclusive, between 1988 and 1997. NCDS and BCS regressions include indicators for ethnicity, and the NCDS regressions also include an age indicator. The PSID regressions include a set of age and year indicators.

J Polit Econ. 2008; 116(3): 499–532.
http://www.journals.uchicago.edu/doi/abs/10.1086/589524



Strong evidence for a 2-3% correlation.

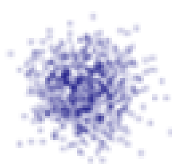
- this doesn't mean being tall *causes* you earn more (though it could)

Correlation

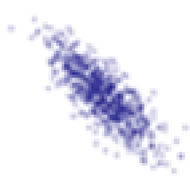
Which of the follow scatter plots shows data with the most negative correlation r?

1.

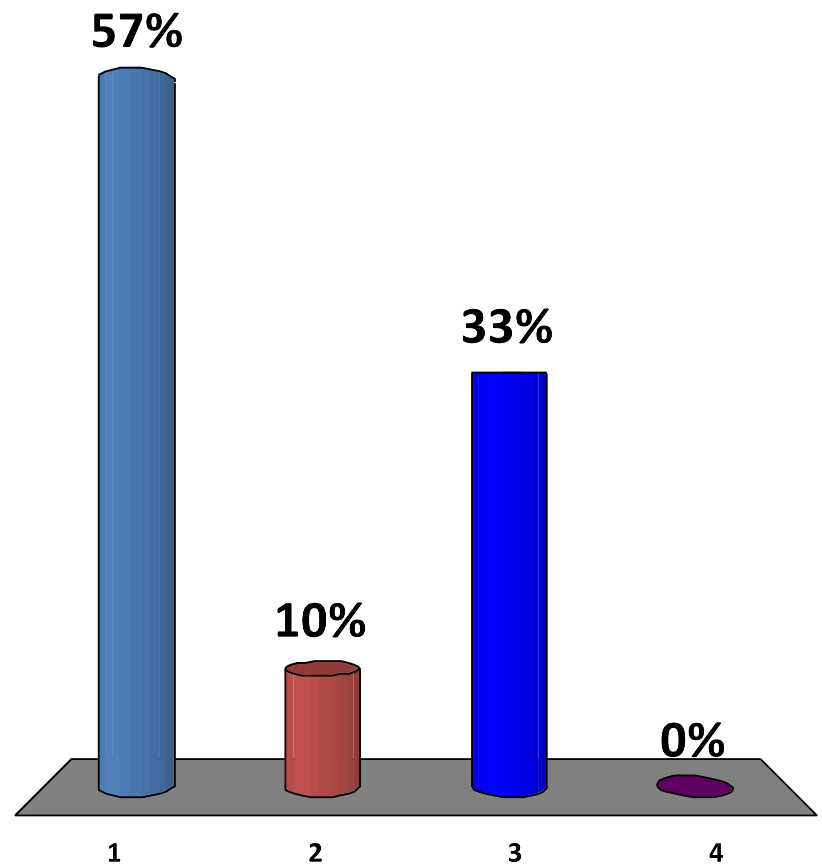
2.



3.



4.



Acceptance Sampling

Situation: large batches of items are produced. We want to sample a small proportion of each batch to check that the proportion of defective items is sufficiently low.

One-stage sampling plans

Sample n items
$X =$ number of defective items in the sample

Reject batch if $X > c$, accept if $X \leq c$

How do we choose n and c?

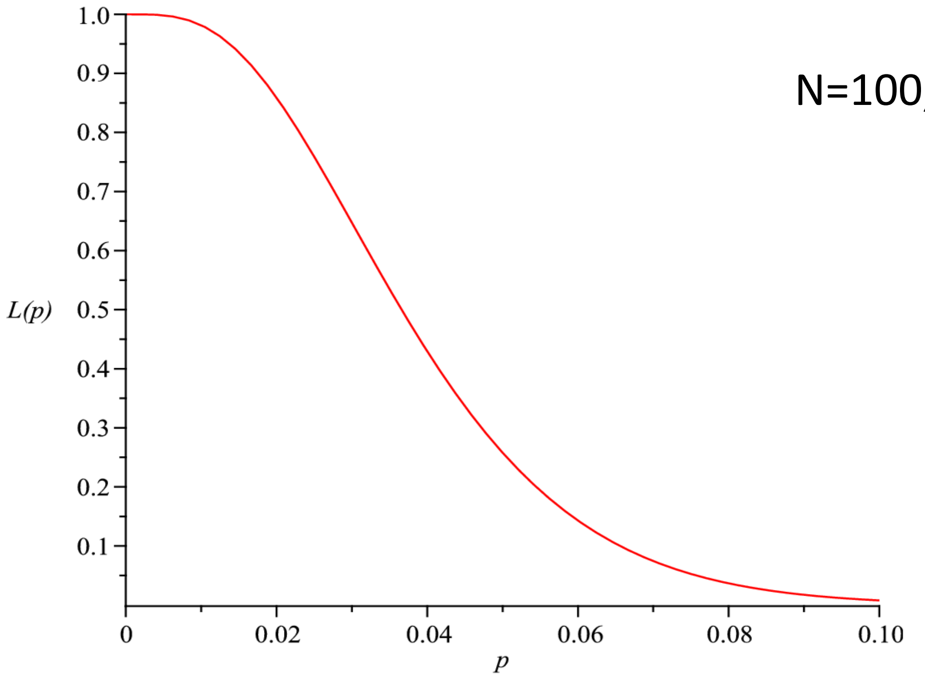
Define $p =$ proportion of defective items in the batch (typically small).

Then $X \sim B(n, p)$ if the population the samples are drawn from is large.

Operating characteristic (OC): probability of accepting the batch

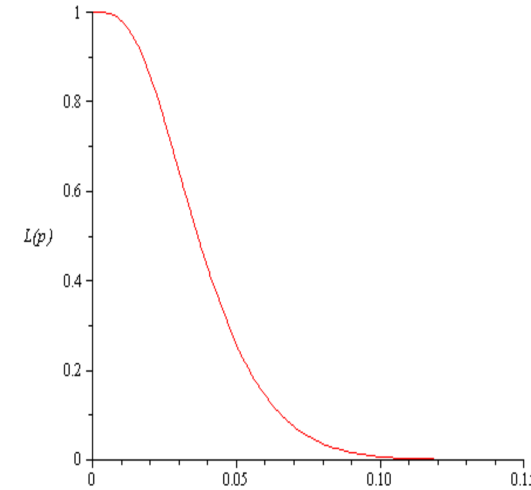
$$L(p) = P(X \leq c)$$

$$= \sum_{k=0}^{c} P(X = k) = \sum_{k=0}^{c} C_k^n \; p^k (1-p)^{n-k}$$

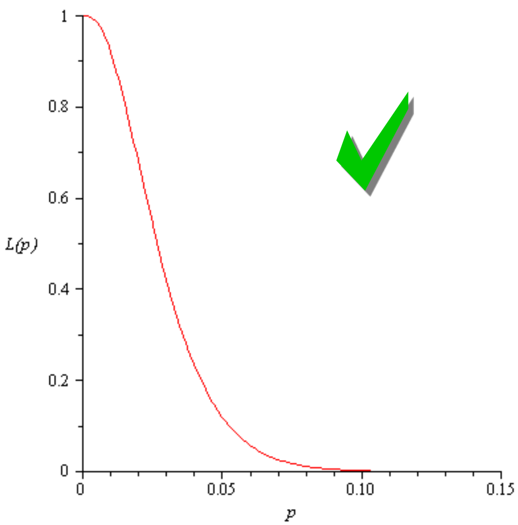


N=100, c=3

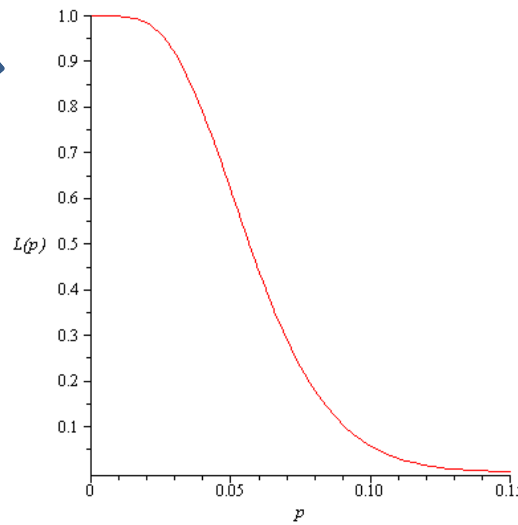
Testing 100 samples and rejecting if more than 3 are faulty gives the OC curve $L(p)$ on the right. Which of the following is the curve for testing 100 samples and rejecting if more than 2 are faulty?



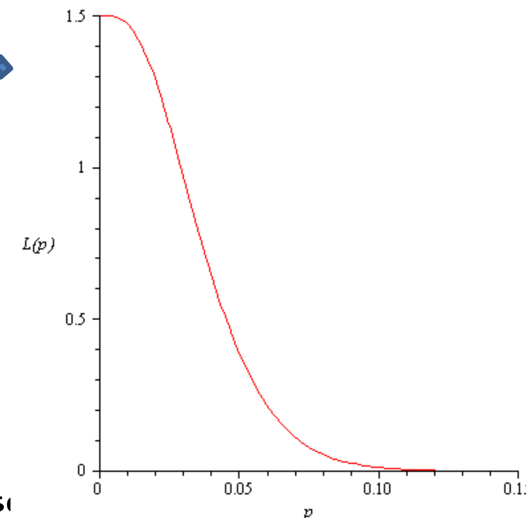
1.



2.

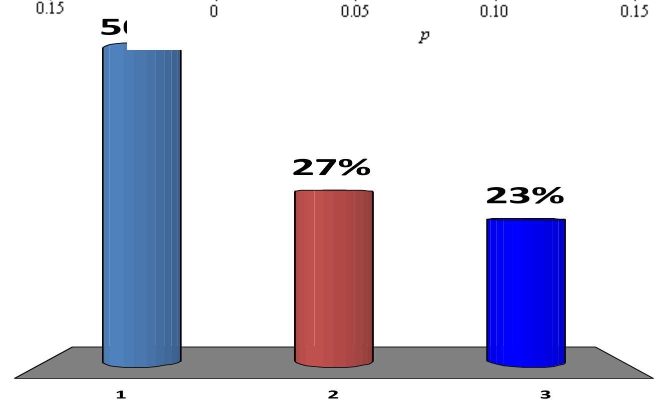


3.



Rejecting more than 2, rather than more than 3 makes it *more likely* to reject the batch (for any p). $\Rightarrow P(\text{rejecting})$ is higher.

$\Rightarrow P(\text{accepting})$ is lower, $\Rightarrow L(p)$ lower



For standard acceptance sampling, Producer and Consumer must decide on the following:

Acceptable quality level: p_1

(consumer happy, want to accept with high probability)

Unacceptable quality level: p_2

(consumer unhappy, want to reject with high probability)

Ideally:

- always accept batch if $p \leq p_1$
- always reject batch if $p \geq p_2$

i.e. $L(P \leq p_1) = 1$ and $L(p \geq p_2) = 0$

- but can't do this without inspecting the entire batch

Use a sampling scheme

Want to minimize:

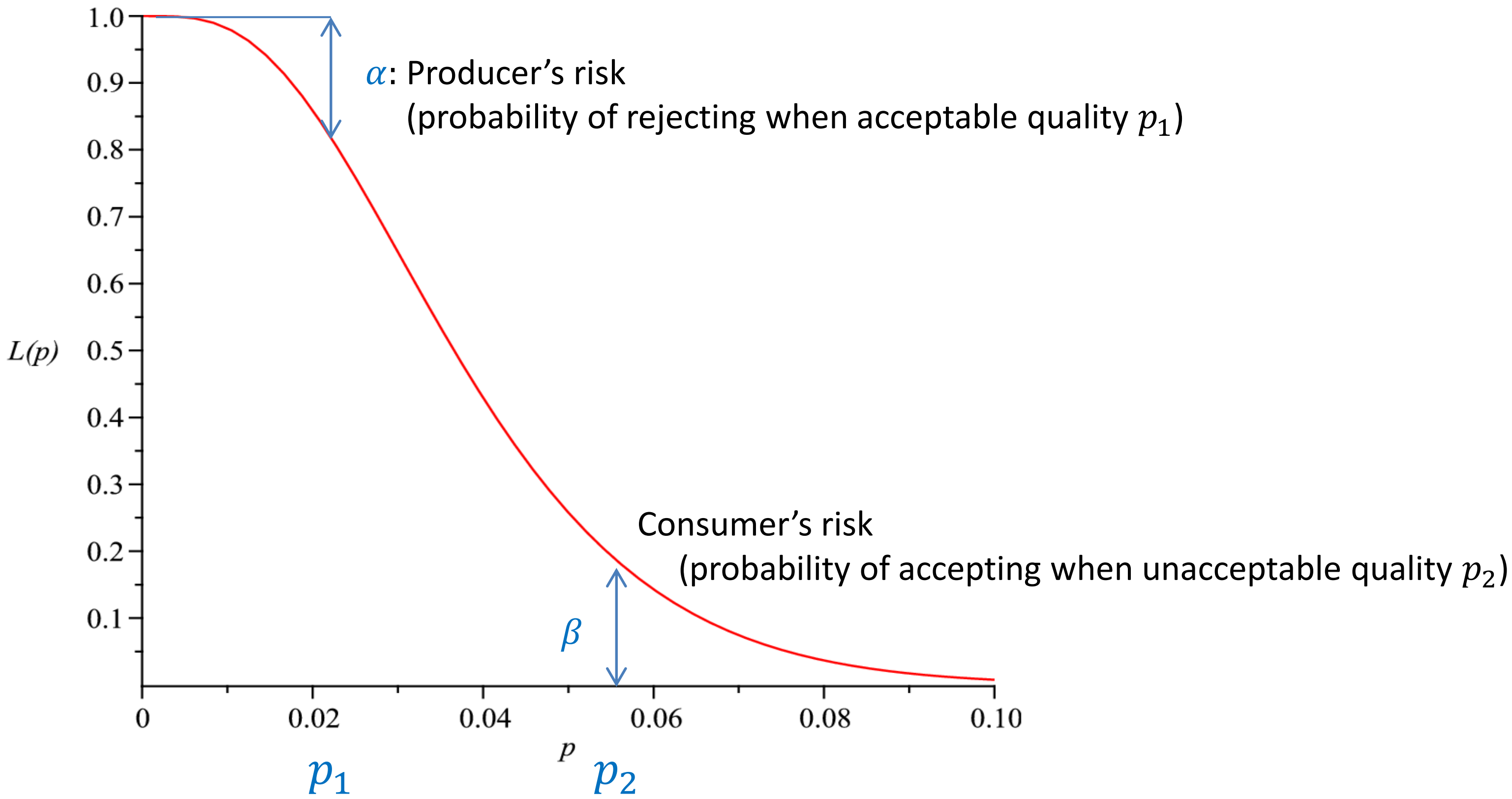
Producer's Risk: reject a batch that has acceptable quality

$$\alpha = P(\text{Reject batch when } p = p_1) \qquad = 1 - L(p_1)$$

Consumer's Risk: accept a batch that has unacceptable quality

$$\beta = P(\text{Accept batch when } p = p_2) \qquad = L(p_2)$$

Operating characteristic curve $L(p)$: probability of accepting the batch



If consumer and producer agree on α, β, p_1, p_2 - can then calculate n and c.

Acceptance Sampling Tables: give n, c for $\alpha = \beta = 0.1$ and $\alpha = \beta = 0.05$

Acceptance Sampling

Sampling plan to achieve Producer's Risk = α, Consumer's Risk = β, acceptable quality level = p_1, unacceptable quality level = p_2.

$\alpha = 0.05$, $\beta = 0.05$

n = sample size, c = max. no. defectives in sample

	p_1																			
	0.005		0.010		0.015		0.020		0.025		0.030		0.035		0.040		0.045		0.050	
p_2	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.010	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.015	1044	9	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.020	523	5	1642	23	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.025	364	4	823	13	2090	40	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.030	257	3	521	9	1093	23	2630	64	*	*	*	*	*	*	*	*	*	*	*	*
0.035	220	3	373	7	656	15	1297	34	*	*	*	*	*	*	*	*	*	*	*	*
0.040	156	2	261	5	452	11	781	22	1500	47	*	*	*	*	*	*	*	*	*	*
0.045	138	2	231	5	346	9	536	16	899	30	1726	63	*	*	*	*	*	*	*	*
0.050	124	2	181	4	260	7	386	12	624	22	1042	40	1947	81	*	*	*	*	*	*
0.055	113	2	164	4	213	6	305	10	460	17	693	28	1170	51	2183	102	*	*	*	*
0.060	103	2	127	3	173	5	259	9	361	14	519	22	787	36	1293	63	2394	124	*	*
0.065	95	2	117	3	159	5	200	7	277	11	388	17	568	27	864	44	1412	76	2603	148
0.070	66	1	109	3	129	4	185	7	239	10	309	14	428	21	625	33	930	52	1528	90
0.075	62	1	82	2	120	4	155	6	206	9	256	12	352	18	476	26	674	39	1018	62
0.080	58	1	77	2	112	4	129	5	178	8	224	11	285	15	374	21	518	31	729	46
0.085	54	1	72	2	89	3	121	5	152	7	182	9	240	13	310	18	406	25	554	36
0.090	51	1	68	2	84	3	115	5	129	6	158	8	199	11	253	15	332	21	434	29
0.095	49	1	65	2	80	3	94	4	122	6	149	8	175	10	227	14	277	18	351	24
0.100	46	1	61	2	76	3	89	4	103	5	129	7	154	9	191	12	239	16	298	21

D R Robinson (1997)

Acceptance Sampling

Sampling plan to achieve Producer's Risk = α, Consumer's Risk = β, acceptable quality level = p_1, unacceptable quality level = p_2.

$\alpha = 0.1$, $\beta = 0.1$ **n = sample size, c = max. no. defectives in sample**

	p_1																			
	0.005		0.010		0.015		0.020		0.025		0.030		0.035		0.040		0.045		0.050	
p_2	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.010	2010	14	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.015	617	5	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.020	333	3	945	13	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.025	212	2	518	8	1310	25	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.030	176	2	308	5	629	13	1609	39	*	*	*	*	*	*	*	*	*	*	*	*
0.035	151	2	227	4	404	9	770	20	1909	56	*	*	*	*	*	*	*	*	*	*
0.040	96	1	166	3	292	7	471	13	926	29	2235	77	*	*	*	*	*	*	*	*
0.045	85	1	147	3	204	5	340	10	547	18	1047	38	2531	100	*	*	*	*	*	*
0.050	77	1	105	2	158	4	258	8	377	13	628	24	1204	50	2828	126	*	*	*	*
0.055	70	1	96	2	144	4	190	6	278	10	426	17	713	31	1331	62	*	*	*	*
0.060	64	1	88	2	110	3	153	5	215	8	313	13	485	22	782	38	1452	75	*	*
0.065	59	1	81	2	101	3	121	4	179	7	235	10	343	16	517	26	857	46	1586	90
0.070	55	1	75	2	94	3	113	4	149	6	201	9	268	13	383	20	575	32	935	55
0.075	51	1	51	1	70	2	105	4	122	5	155	7	219	11	297	16	418	24	610	37
0.080	48	1	48	1	65	2	82	3	98	4	130	6	175	9	234	13	321	19	447	28
0.085	45	1	45	1	61	2	77	3	93	4	122	6	151	8	193	11	248	15	342	22
0.090	42	1	42	1	58	2	73	3	87	4	101	5	129	7	156	9	208	13	272	18
0.095	40	1	40	1	55	2	55	2	69	3	96	5	109	6	135	8	172	11	221	15
0.100	38	1	38	1	52	2	52	2	65	3	78	4	104	6	116	7	152	10	187	13

Example

In planning an acceptance sampling scheme, the Producer and Consumer have agreed that the acceptable quality level is 2% defectives and the unacceptable level is 6%. Each is prepared to take a 10% risk. What sample size is required and under what circumstances should the batch be rejected?

Answer

$$\alpha = \beta = 0.1, p_1 = 0.02, p_2 = 0.06$$

$$\Rightarrow n = 153, c = 5$$

Should sample 153 items and reject if the number of defective items is greater than 5.

Mathematical Sciences **STATISTICAL TABLES**

Acceptance Sampling

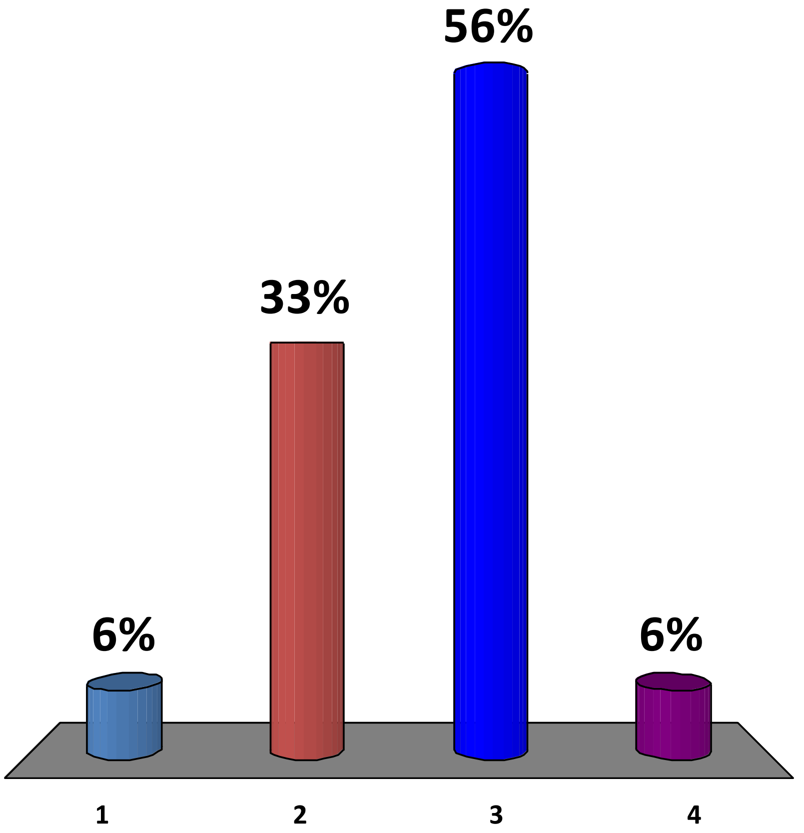
Sampling plan to achieve Producer's Risk = α, Consumer's Risk = β, acceptable quality lev

$\alpha = 0.1, \beta = 0.1$

	p_1											
	0.005		0.010		0.015		0.020		0.025		0.030	
p_2	n	c	n	c	n	c	n	c	n	c	n	c
0.010	2010	14	*	*	*	*	*	*	*	*	*	*
0.015	617	5	*	*	*	*	*	*	*	*	*	*
0.020	333	3	945	13	*	*	*	*	*	*	*	*
0.025	212	2	518	8	1310	25	*	*	*	*	*	*
0.030	176	2	308	5	629	13	1609	39	*	*	*	*
0.035	151	2	227	4	404	9	770	20	1909	56	*	*
0.040	96	1	166	3	292	7	471	13	926	29	2235	77
0.045	85	1	147	3	204	5	340	10	547	18	1047	38
0.050	77	1	105	2	158	4	258	8	377	13	628	24
0.055	70	1	96	2	144	4	190	6	278	10	426	17
0.060	64	1	88	2	110	3	153	5	215	8	313	13
0.065	59	1	81	2	101	3	121	4	179	7	235	10
0.070	55	1	75	2	94	3	113	4	149	6	201	9
0.075	51	1	51	1	70	2	105	4	122	5	155	7
0.080	48	1	48	1	65	2	82	3	98	4	130	6
0.085	45	1	45	1	61	2	77	3	93	4	122	6
0.090	42	1	42	1	58	2	73	3	87	4	101	5
0.095	40	1	40	1	55	2	55	2	69	3	96	5
0.100	38	1	38	1	52	2	52	2	65	3	78	4

In planning an acceptance sampling scheme, the Producer and Consumer have agreed that the acceptable quality level is 1% defectives and the unacceptable level is 3%. Each is prepared to take a 5% risk. What is the best plan?

- sample 308 items and reject if the number of defective items is greater than 5
- sample 308 items and reject if the number of defective items is 5 or more
- sample 521 items and reject if the number of defective items is 9 or more
- sample 521 items and reject if the number of defective items is 10 or more



In planning an acceptance sampling scheme, the Producer and Consumer have agreed that the acceptable quality level is 1% defectives and the unacceptable level is 3%. Each is prepared to take a 5% risk.

What is the best plan?

Mathematical Sciences	STATISTICAL TABLES	University of Sussex

Acceptance Sampling

Sampling plan to achieve Producer's Risk = α, Consumer's Risk = β, acceptable quality level = p_1, unacceptable quality level = p_2.

$\alpha = 0.05, \beta = 0.05$

n = sample size, c = max. no. defectives in sample

	p_1																			
	0.005		0.010		0.015		0.020		0.025		0.030		0.035		0.040		0.045		0.050	
p_2	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.010	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.015	1044	9	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.020	523	5	1642	23	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.025	364	4	823	13	2090	40	*	*	*	*	*	*	*	*	*	*	*	*	*	*
0.030	257	3	521	9	1093	23	2630	64	*	*	*	*	*	*	*	*	*	*	*	*
0.035	220	3	373	7	656	15	1297	34	*	*	*	*	*	*	*	*	*	*	*	*
0.040	156	2	261	5	452	11	781	22	1500	47	*	*	*	*	*	*	*	*	*	*
0.045	138	2	231	5	346	9	536	16	899	30	1726	63	*	*	*	*	*	*	*	*
0.050	124	2	181	4	260	7	386	12	624	22	1042	40	1947	81	*	*	*	*	*	*
0.055	113	2	164	4	213	6	305	10	460	17	693	28	1170	51	2183	102	*	*	*	*
0.060	103	2	127	3	173	5	259	9	361	14	519	22	787	36	1293	63	2394	124	*	*
0.065	95	2	117	3	159	5	200	7	277	11	388	17	568	27	864	44	1412	76	2603	148
0.070	66	1	109	3	129	4	185	7	239	10	309	14	428	21	625	33	930	52	1528	90
0.075	62	1	82	2	120	4	155	6	206	9	256	12	352	18	476	26	674	39	1018	62
0.080	58	1	77	2	112	4	129	5	178	8	224	11	285	15	374	21	518	31	729	46
0.085	54	1	72	2	89	3	121	5	152	7	182	9	240	13	310	18	406	25	554	36
0.090	51	1	68	2	84	3	115	5	129	6	158	8	199	11	253	15	332	21	434	29
0.095	49	1	65	2	80	3	94	4	122	6	149	8	175	10	227	14	277	18	351	24
0.100	46	1	61	2	76	3	89	4	103	5	129	7	154	9	191	12	239	16	298	21

D R Robinson (1997)

Sample 521 and reject if more than 9 (i.e. 10 or more)

Example – calculating the risks

It has been decided to sample 100 items at random from each large batch and to reject the batch if more than 2 defectives are found. The acceptable quality level is 1% and the unacceptable quality level is 5%.
Find the Producer's and Consumer's risks.

Answer

$n = 100, c = 2, p_1 = 0.01, p_2 = 0.05$

1. For the Producer's Risk: want probability of reject batch when $p = p_1 = 0.01$

$X \sim B(100{,}0.01)$

$P(\text{Reject batch when } p = 0.01) = 1 - P(\text{accept batch}) = 1 - L(0.01)$

$= 1 - P(X = 0) - P(X = 1) - P(X = 2)$

$= 1 - C_0^{100} 0.01^0 \times 0.99^{100} - C_1^{100} 0.01 \times 0.99^{99} - C_2^{100} 0.01^2 \times 0.99^{98}$

$=$ 1 - 0.3660 - 0.3697 - 0.1849 = 0.079.

Example – calculating the risks

It has been decided to sample 100 items at random from each large batch and to reject the batch if more than 2 defectives are found. The acceptable quality level is 1% and the unacceptable quality level is 5%.
Find the Producer's and Consumer's risks.

Answer

$n = 100, c = 2, p_1 = 0.01, p_2 = 0.05$

2. For the Consumer's Risk: want probability of accepting batch when $p = p_2 = 0.05$

$X \sim B(100,0.05)$

$P(\text{Accept when } p = 0.05) = L(0.05)$

$= P(X = 0) + P(X = 1) + P(X = 2)$

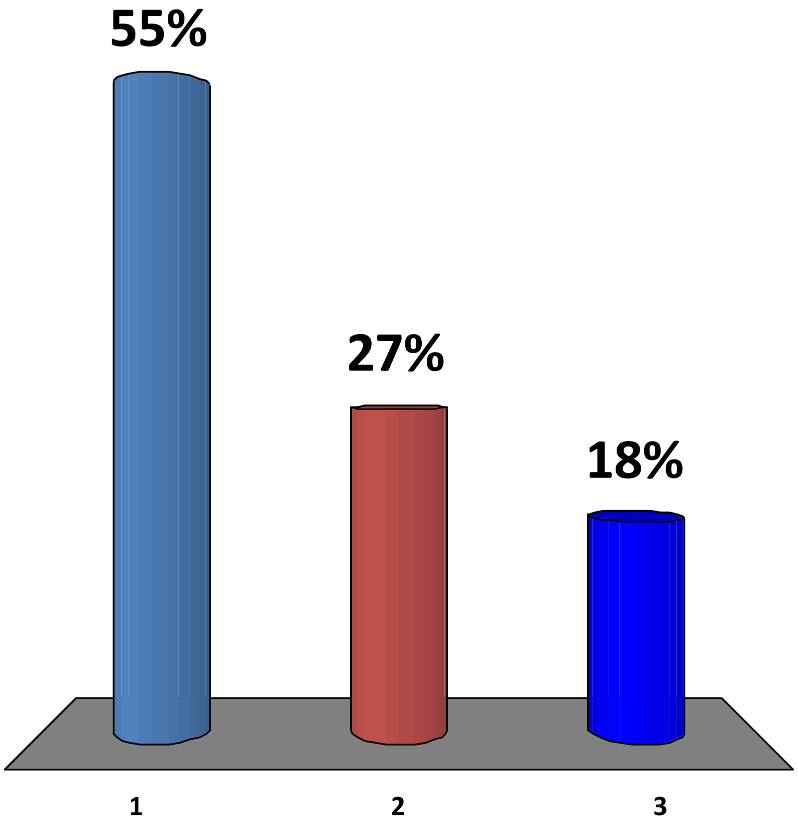
$= C_0^{100} 0.05^0 \times 0.95^{100} + C_1^{100} 0.05 \times 0.95^{99} + C_2^{100} 0.05^2 \times 0.95^{98}$

$= 0.118$

It has been decided to sample 100 items at random from each large batch and to reject the batch if more than 2 defectives are found. The acceptable quality level is 1% and the unacceptable quality level is 5%.

Which of the following would increase the Consumer's Risk?

1. Increasing the acceptable quality level to 2%
2. Decreasing the unacceptable quality level to 4%
3. Rejecting if more than 1 defectives are found



It has been decided to sample 100 items at random from each large batch and to reject the batch if more than 2 defectives are found. The acceptable quality level is 1% and the unacceptable quality level is 5%.

Which of the following would increase the Consumer's Risk?

Increasing the acceptable quality level

NO – Consumer's Risk depends on the unacceptable quality level

Decreasing the unacceptable quality level

YES –e.g. then more likely to accept when the
defect probability is $p = 0.04$ compared to $p = 0.05$

Rejecting if more than 1 defectives are found

NO – more likely to get 1 or more, so less likely to accept batch
$\Rightarrow$ lower Consumer's Risk

Two-stage sampling plan

Idea: test some, reject if clearly bad, accept if clearly good, if not clear investigate further

1. Sample n_1 items, $X_1 =$ number of defectives in the sample

2. Accept batch if $X_1 \leq c_1$, reject if $X_1 > c_2$ (where $c_2 > c_1$)

3. If $c_1 < X_1 \leq c_2$, sample a further n_2 items;
   let $X_2 =$ number of defectives in 2nd sample

4. Accept batch if $X_2 \leq c_3$, otherwise reject batch.

Advantage: can require fewer samples than single-stage plan (for similar $L(p)$)

Distadvantage: more complicated, need to choose n_1, n_2, c_1, c_2, c_3

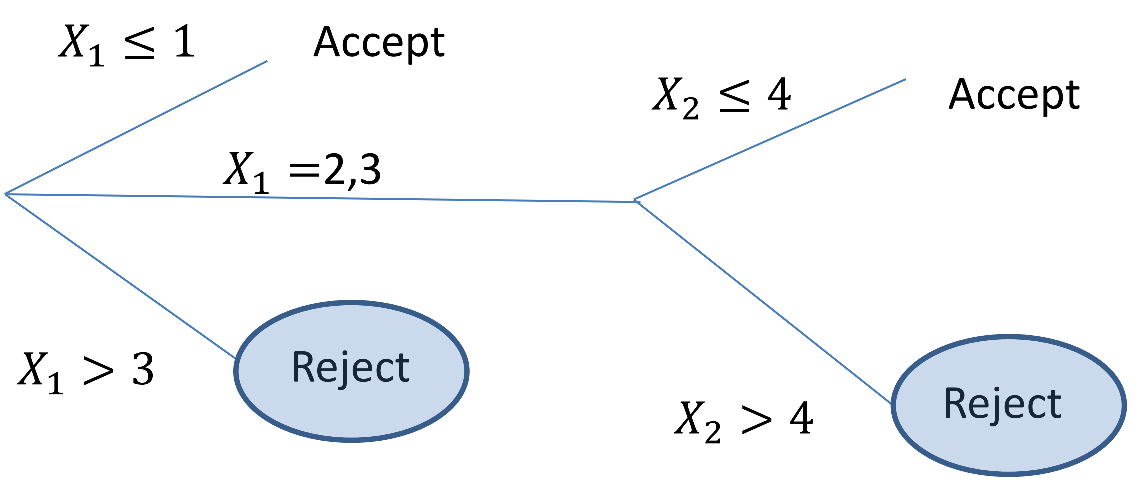
Example

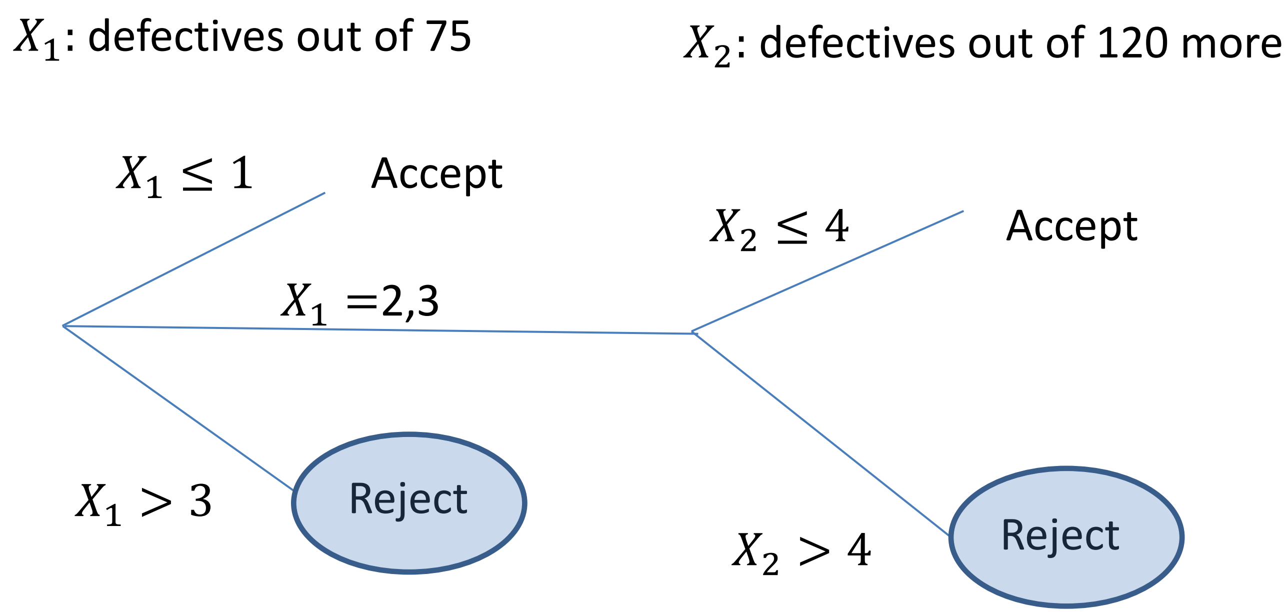
A two-stage sampling plan for a quality control procedure is as follows: Sample 75 items, accept if less than 2 defectives, reject if more than 3 defectives; otherwise sample 120 more and reject if more than 4 defectives in the new batch

Find the probability that a batch is rejected under this plan if the probability p of any particular item being faulty is $p = 0.02$.

Answer

Let X_1 be number faulty in the first batch, X_2 be number faulty in second batch (if taken)





$$P(\text{reject}) = P(\text{reject in first stage}) + P(X_1 = 2{,}3)P(\text{reject in second stage})$$

$$= P(X_1 > 3) + [P(X_1 = 2) + P(X_1 = 3)]P(X_2 > 4)$$

$$= 1 - \sum_{n=0}^{3} P(X_1 = n) + [P(X_1 = 2) + P(X_1 = 3)][1 - \sum_{n=0}^{4} P(X_2 = n)]$$

$$= 1 - (1 - 0.01)^{75} - \mathrm{C}_1^{75} 0.01(1 - 0.01)^{74} - \mathrm{C}_2^{75} 0.01^2 (1 - 0.01)^{73} + \ldots$$

$$\approx 0.099$$

Example (as before)

In planning an acceptance sampling scheme, the Producer and Consumer have agreed that the acceptable quality level is 2% defectives and the unacceptable level is 6%. Each is prepared to take a 10% risk. What sample size is required and under what circumstances should the batch be rejected?

$\alpha = \beta = 0.1, p_1 = 0.02, p_2 = 0.06$

Answer: single stage plan $n = 153, c = 5$

Sample 153 items and reject if the number of defective items is greater than 5.

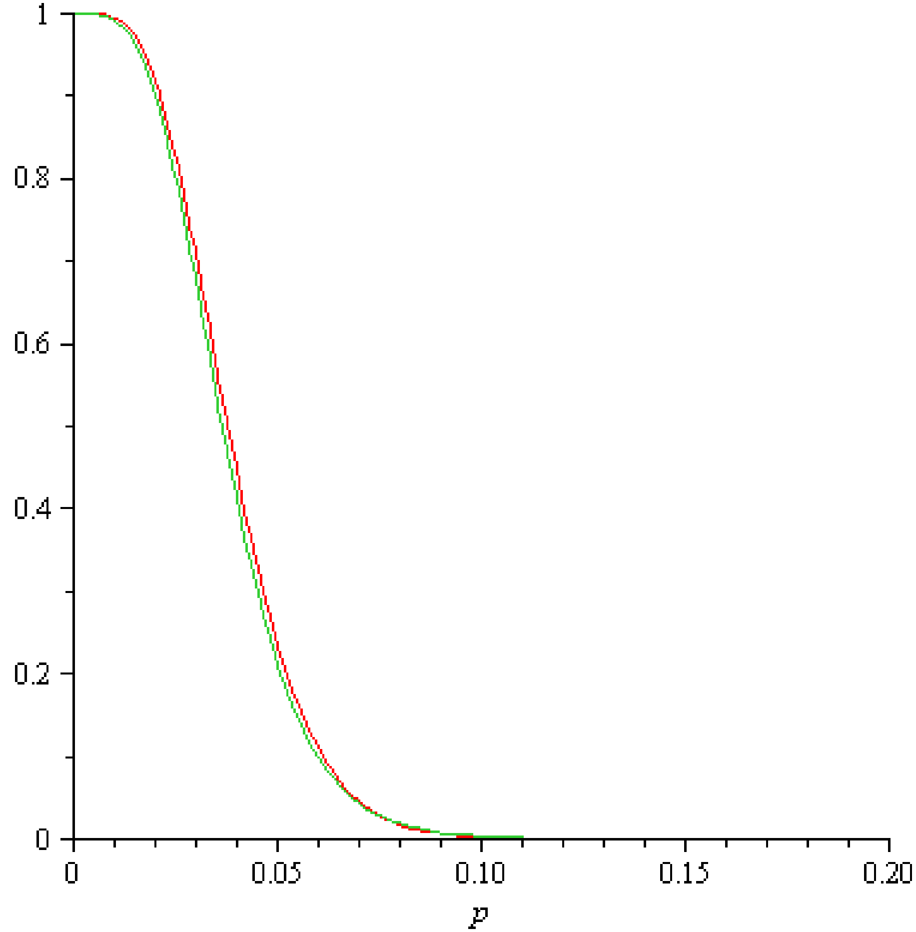
- Always take 153 samples

Alternative answer: two-stage plan, as last example

$n_1 = 75, n_2 = 120, c_1 = 1, \mathrm{c}_2 = 3, \mathrm{c}_3 = 4$

- Sometimes takes only 75 samples, sometimes 120+75=195
- Mean number: 75 to 132 (depending on p); more efficient!

Two-stage plan can have very similar OC curve, but require fewer samples



BUT: - not obvious how to choose n_1, n_2, c_1, c_2, c_3; example not optimal

- less parallelizable (e.g. might care if testing is cheap but takes a long time)

Variation: better to include first sample with second sample for final decision